

日本国特許庁
JAPAN PATENT OFFICE

340300993

07.20.04

別紙添付の書類に記載されている事項は下記の出願書類に記載されて
いる事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed
with this Office.

出願年月日 2003年10月15日
Date of Application:

出願番号 特願2003-354557
Application Number:
[ST. 10/C]: [JP2003-354557]

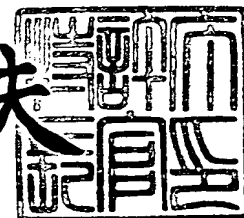
願人 株式会社日立製作所
Applicant(s):

CERTIFIED COPY OF
PRIORITY DOCUMENT

2004年 6月21日

特許庁長官
Commissioner,
Japan Patent Office

今井康夫



出証番号 出証特2004-3053582

【書類名】 特許願
【整理番号】 340300993
【あて先】 特許庁長官殿
【国際特許分類】 G06F 03/06
【発明者】
 【住所又は居所】 神奈川県小田原市中里 3 2 2 番 2 号 株式会社日立製作所 R A I
 D システム事業部内
 【氏名】 佐藤 元
【発明者】
 【住所又は居所】 神奈川県小田原市中里 3 2 2 番 2 号 株式会社日立製作所 R A I
 D システム事業部内
 【氏名】 本間 繁雄
【特許出願人】
 【識別番号】 000005108
 【氏名又は名称】 株式会社日立製作所
【代理人】
 【識別番号】 100095371
 【弁理士】
 【氏名又は名称】 上村 輝之
【選任した代理人】
 【識別番号】 100089277
 【弁理士】
 【氏名又は名称】 宮川 長夫
【選任した代理人】
 【識別番号】 100104891
 【弁理士】
 【氏名又は名称】 中村 猛
【手数料の表示】
 【予納台帳番号】 043557
 【納付金額】 21,000円
【提出物件の目録】
 【物件名】 特許請求の範囲 1
 【物件名】 明細書 1
 【物件名】 図面 1
 【物件名】 要約書 1
 【包括委任状番号】 0110323

【書類名】 特許請求の範囲**【請求項 1】**

上位装置と通信可能に接続され得るディスクアレイ装置において、

このディスクアレイ装置の全体の制御を行うディスクアレイ制御部と、

前記上位装置とのデータ転送を制御する上位側転送制御部と、

少なくとも、1つのパリティグループを構成する複数のデータディスクドライブと、1以上のスペアディスクドライブとを有するディスクアレイであって、前記1つのパリティグループは前記複数のデータディスクドライブの記憶領域に渡って形成される多数のデータストライプを有し、前記多数のデータストライプは前記データストライプの2以上のセットに分けることができる、前記ディスクアレイと、

前記上位装置及び前記ディスクアレイの間で転送されるデータの一時記憶に用いられるキャッシュメモリと、

前記ディスクアレイとのデータ転送を制御する下位側転送制御部とを備え、

前記ディスクアレイ制御部が、

前記データディスクドライブ毎の故障発生の可能性を推定する推定部と、

前記複数のデータディスクドライブの内から、前記推定された故障発生の可能性に応じて2以上のデータディスクドライブを、分割データコピーの対象として選び、前記選ばれた2以上のデータディスクドライブの各々から一つずつ分割記憶領域を分割することにより2以上の分割記憶領域を選び、前記選ばれた2以上の分割記憶領域は、前記パリティグループ内の前記データストライプの異なるセットにそれぞれ属し、そして、前記選ばれた2以上の分割記憶領域のデータを前記1以上のスペアディスクドライブへコピーするように前記下位側転送制御部及び前記キャッシュメモリを制御する分割データコピー部とを有するディスクアレイ装置。

【請求項 2】

請求項1記載のディスクアレイ装置において、

前記ディスクアレイ制御部が、

前記複数のデータディスクドライブの中から、前記推定された故障発生の可能性に応じて1つのデータディスクドライブを、ダイナミックスペアリングの対象として選び、前記選ばれた1つのデータディスクドライブから、前記分割データコピー部によるコピーが未だ行われていない残りの分割記憶領域を選び、そして、前記選ばれた残りの分割記憶領域のデータを前記スペアディスクドライブへコピーするように前記下位側転送制御部及び前記キャッシュメモリを制御するダイナミックスペアリング部を更に有するディスクアレイ装置。

【請求項 3】

請求項2記載のディスクアレイ装置において、

前記分割データコピー部は、前記複数のデータディスクドライブの内の第1のデータディスクドライブの前記推定された故障発生の可能性が第1のレベルに達した場合、少なくとも、前記第1のデータディスクドライブと他の第2のデータディスクドライブとを、前記分割データコピーの対象として選択し、

前記ダイナミックスペアリング部は、前記第1のデータディスクドライブの前記推定された故障発生の可能性が前記第1のレベルより高い第2のレベルに達した場合、前記第1のデータディスクドライブを、前記ダイナミックスペアリングの対象として選択するディスクアレイ装置。

【請求項 4】

請求項1記載のディスクアレイ装置において、

前記分割データコピー部は、前記複数のデータディスクドライブの内の第1のデータディスクドライブの前記推定された故障発生の可能性が所定のレベルに達した場合、前記第1のデータディスクドライブと、前記推定された故障発生の可能性が前記第1のデータディスクドライブに次いで高い第2のデータディスクドライブとを、前記分割データコピー

の対象として選択するディスクアレイ装置。

【請求項 5】

請求項 1 記載のディスクアレイ装置において、

前記分割データコピー部は、前記選ばれた 2 以上の分割記憶領域のデータをコピーする過程で、前記選ばれた 2 以上の分割記憶領域から同時にデータを読み出すように前記下位側転送制御部及び前記キャッシュメモリを制御するディスクアレイ装置。

【請求項 6】

請求項 1 記載のディスクアレイ装置において、

前記ディスクアレイ制御部が、

前記選ばれた 2 以上の分割記憶領域のデータの前記 1 以上のスเปアディスクドライブへのコピーが開始された後、前記上位装置からの前記選ばれた 2 以上の分割記憶領域への新たなデータの書き込み要求を前記上位側データ転送部から受けた場合、前記上位装置から受けた前記新たなデータを、前記選ばれた 2 以上の分割記憶領域へ書き込むと共に、前記スเปアディスクドライブにも書き込むように前記下位側転送制御部及び前記キャッシュメモリを制御するスเปアデータ最新化部を更に有するディスクアレイ装置。

【請求項 7】

請求項 1 記載のディスクアレイ装置において、

前記分割データコピー部が、前記複数のデータディスクドライブから前記分割データコピーの対象として第 1 及び第 2 のデータディスクドライブを選び、前記第 1 のデータディスクドライブから前記パリティグループの内の前側のデータストライプのセットに属する第 1 の分割記憶領域を選び、前記第 2 のデータディスクドライブから前記前側のデータストライプのセットに引き続く後側のデータストライプのセットに属する第 2 の分割記憶領域を選び、そして、前記第 1 と第 2 の分割記憶領域のデータを前記スเปアディスクドライブへコピーするように前記下位側転送制御部及び前記キャッシュメモリを制御するディスクアレイ装置。

【請求項 8】

請求項 1 記載のディスクアレイ装置において、

前記分割データコピー部により前記 2 以上のデータディスクドライブから選ばれる前記 2 以上の分割記憶領域が、実質的に等しいサイズをもつディスクアレイ装置。

【請求項 9】

請求項 1 記載のディスクアレイ装置において、

前記分割データコピー部により前記 2 以上のデータディスクドライブから選ばれる前記 2 以上の分割記憶領域が、前記 2 以上のデータディスクドライブの前記推定された故障発生の可能性に応じて異なるサイズを有するディスクアレイ装置。

【請求項 10】

請求項 1 記載のディスクアレイ装置において、

前記推定部が、前記データディスクドライブ毎のエラー発生履歴を記憶し、前記記憶されたエラー発生履歴に基づいて前記データディスクドライブ毎に前記故障発生の可能性を推定するディスクアレイ装置。

【請求項 11】

上位装置と通信可能に接続され得るディスクアレイ装置であって、

このディスクアレイ装置の全体の制御を行うディスクアレイ制御部と、前記上位装置とのデータ転送を制御する上位側転送制御部と、

少なくとも、1つのパリティグループを構成する複数のデータディスクドライブと、1以上のスเปアディスクドライブとを有するディスクアレイであって、前記 1つのパリティグループは前記複数のデータディスクドライブの記憶領域に渡って形成される多数のデータストライプを有し、前記多数のデータストライプは前記データストライプの 2 以上のセットに分けることができる、前記ディスクアレイと、

前記上位装置及び前記ディスクアレイの間で転送されるデータの一時記憶に用いられるキャッシュメモリと、

前記ディスクアレイとのデータ転送を制御する下位側転送制御部とを備えたものにおいて、

前記ディスクアレイ制御部が前記スペアディスクドライブを用いて前記データディスクドライブ内のデータをスペアするための方法において、

前記データディスクドライブ毎の故障発生の可能性を推定するステップと、

前記複数のデータディスクドライブの内から、前記推定された故障発生の可能性に応じて、2以上のデータディスクドライブを、分割データコピーの対象として選ぶステップと、

前記選ばれた2以上のデータディスクドライブの各々から一つずつ分割記憶領域を分割することにより2以上の分割記憶領域を選ぶステップであって、前記選ばれた2以上の分割記憶領域が、前記パリティグループ内の異なるデータストライプのセットにそれぞれ属するようになったステップと、

前記選ばれた2以上の分割記憶領域のデータを前記1以上のスペアディスクドライブへコピーするように前記下位側転送制御部及び前記キャッシュメモリを制御することで分割データコピーを行うステップと

を有することを特徴とする方法。

【請求項12】

請求項11記載の方法において、

前記複数のデータディスクドライブの中から、前記推定された故障発生の可能性に応じて1つのデータディスクドライブを、ダイナミックスペアリングの対象として選ぶステップと、

前記選ばれた1つのデータディスクドライブから、前記スペアディスクへのコピーが未だ行われていない残りの分割記憶領域を選び、前記選ばれた残りの分割記憶領域のデータを前記スペアディスクドライブへコピーするように前記下位側転送制御部及び前記キャッシュメモリを制御することでダイナミックスペアリングを行うステップを更に有する方法。

【請求項13】

請求項12記載の方法において、

前記2以上のデータディスクドライブを選ぶステップは、前記複数のデータディスクドライブの内の第1のデータディスクドライブの前記推定された故障発生の可能性が第1のレベルに達した場合に行なわれて、そこで、少なくとも前記第1のデータディスクドライブと他の第2のデータディスクドライブとが、前記分割データコピーの対象として選択され、

前記1つのデータディスクドライブを選ぶステップは、前記第1のデータディスクドライブの前記推定された故障発生の可能性が前記第1のレベルより高い第2のレベルに達した場合に行なわれて、そこで、前記第1のデータディスクドライブが、前記ダイナミックスペアリングの対象として選択される方法。

【請求項14】

請求項11記載の方法において、

前記2以上のデータディスクドライブを選ぶステップは、前記複数のデータディスクドライブの内の第1のデータディスクドライブの前記推定された故障発生の可能性が所定のレベルに達した場合に行なわれて、そこで、前記第1のデータディスクドライブと、前記推定された故障発生の可能性が前記第1のデータディスクドライブに次いで高い第2のデータディスクドライブとが、前記分割データコピーの対象として選択される方法。

【請求項15】

請求項11記載の方法において、

前記制御するステップでは、前記選ばれた2以上の分割記憶領域から同時にデータが読み出されよう前記下位側転送制御部及び前記キャッシュメモリを制御する方法。

【請求項16】

請求項11記載の方法において、

前記分割データコピーを行うステップが開始された後、前記上位装置からの前記選ばれた2以上の分割記憶領域への新たなデータの書き込み要求を前記上位側データ転送部から受けた場合、前記上位装置から受けた前記新たなデータを、前記選ばれた2以上の分割記憶領域へ書き込むと共に前記スペアディスクドライブにも書き込むように、前記下位側転送制御部及び前記キャッシュメモリを制御することにより、前記スペアディスク内のデータを最新化するステップを更に有する方法。

【請求項17】

請求項11記載の方法において、

前記2以上のデータディスクドライブを選ぶステップでは、前記複数のデータディスクドライブから第1及び第2のデータディスクドライブが選ばれ、

前記2以上の分割記憶領域を選ぶステップでは、前記第1のデータディスクドライブから前記パリティグループ内の前側のデータストライプのセットに属する第1の分割記憶領域が選ばれ、前記第2のデータディスクドライブから前記前側のデータストライプのセットに引き続く後側のデータストライプのセットに属する第2の分割記憶領域が選ばれる方法。

【請求項18】

請求項11記載の方法において、

前記2以上の分割記憶領域を選ぶステップでは、実質的に等しいサイズを有する2以上の分割記憶領域が選ばれる方法。

【請求項19】

請求項11記載のデータスペアリング方法において、

前記2以上の分割記憶領域を選ぶステップでは、前記2以上のデータディスクドライブについて推定された故障発生の可能性に応じて、異なるサイズを有する2以上の分割記憶領域が選ばれる方法。

【請求項20】

上位装置と通信可能に接続され得るディスクアレイ装置であって、

このディスクアレイ装置の全体の制御を行うディスクアレイ制御部と、前記上位装置とのデータ転送を制御する上位側転送制御部と、

少なくとも、1つのパリティグループを構成する複数のデータディスクドライブと、1以上のスペアディスクドライブとを有するディスクアレイであって、前記1つのパリティグループは前記複数のデータディスクドライブの記憶領域に渡って形成される多数のデータストライプを有し、前記多数のデータストライプは前記データストライプの2以上のセットに分けることができる、前記ディスクアレイと、

前記上位装置及び前記ディスクアレイの間で転送されるデータの一時記憶に用いられるキャッシュメモリと、

前記ディスクアレイとのデータ転送を制御する下位側転送制御部とを備えたものにおける、

前記スペアディスクドライブを用いて前記データディスクドライブ内のデータをスペアする動作を制御するための装置において、

前記データディスクドライブ毎の故障発生の可能性を推定する推定部と、

前記複数のデータディスクドライブの内から、前記推定された故障発生の可能性に応じて2以上のデータディスクドライブを、分割データコピーの対象として選び、前記選ばれた2以上のデータディスクドライブの各々から一つずつ分割記憶領域を分割することにより2以上の分割記憶領域を選ぶものであって、前記選ばれた2以上の分割記憶領域が前記パリティグループ内の異なるデータストライプのセットにそれぞれ属するようにする分割領域選択部と、

前記選ばれた2以上の分割記憶領域のデータを前記スペアディスクドライブへコピーするように前記下位側転送制御部及び前記キャッシュメモリを制御する分割データコピー部とを備えたデータスペアリング制御装置。

【書類名】 明細書

【発明の名称】 スペアディスクドライブをもつディスクアレイ装置及びデータスペアリング方法

【技術分野】**【0001】**

本発明は、RAID機能を有する複数のディスクドライブとスペアディスクドライブを持つディスクアレイ装置において、複数のディスクドライブから1台または複数台のスペアディスクドライブへ並列にデータコピーする機能を有するディスクアレイ装置およびその制御方式に関する。

【背景技術】**【0002】**

従来、ディスクアレイ装置では、RAID機能に加えてスペアディスクドライブを持ち、信頼性を向上させている。このようなディスクアレイ装置ではRAIDを構成するディスクドライブの1台に障害が発生したとき、RAIDを構成している他のディスクドライブからデータを復旧してスペアディスクドライブへコピーすること（コレクションコピー）で、RAIDの縮退動作から通常アクセス状態へ復旧する。しかし、データ復旧の際に縮退動作となるため、信頼性と性能の面で問題がある。そこで、ディスクドライブのエラー履歴を解析して故障の可能性を予測し、その可能性の高いディスクドライブのデータを、故障発生前にスペアディスクドライブへコピーすること（ダイナミックスペアリング）で、より信頼性を高めている（特許文献1、2参照）。

【0003】

【特許文献1】特開平11-345095号公報

【特許文献2】特開平08-249133号公報

【発明の開示】**【発明が解決しようとする課題】****【0004】**

しかしながら、最近のディスクドライブは大容量化し、高負荷環境で使用されることも多く、そのため、上述したコレクションコピーやダイナミックスペアリングにかかる時間が増大している。そのため、次のような多重ディスク故障の発生可能性が高まっている。例えば、ダイナミックスペアリングのためのコピー動作中に、コピー元ディスクドライブで障害が頻発しREAD動作を継続できなくなる場合があり得る。この場合、当該ディスクドライブは閉塞され、このディスクドライブが属するパリティグループ（つまり、ECC（エラーコネクティングコード）グループ）についてデータ復旧動作（コレクションコピー）が行われることになる。しかしながら、このコレクションコピー動作中に、同一パリティグループ中の別のディスクドライブで、部分的なセクタ障害或いはハードディスク閉塞などの障害が発生する場合があり得る。このような多重ディスク故障が発生すると、データは喪失される。

【0005】

従って、本発明の目的は、スペアディスクを持つRAID構成のディスクアレイ装置において、ディスクの多重故障によるデータロストの危険性を低減することにある。

【0006】

本発明の別の目的は、高負荷時にも高速なダイナミックスペアリングが可能なディスクアレイ装置を提供することにある。

【課題を解決するための手段】**【0007】**

本発明の一つの観点に従う、上位装置と通信可能に接続され得るディスクアレイ装置は、このディスクアレイ装置の全体の制御を行うディスクアレイ制御部と、前記上位装置とのデータ転送を制御する上位側転送制御部と、少なくとも、1つのパリティグループを構成する複数のデータディスクドライブと、1以上のスペアディスクドライブとを有するディスクアレイであって、前記1つのパリティグループは前記複数のデータディスクドライブ

ブの記憶領域に渡って形成される多数のデータストライプを有し、前記多数のデータストライプは前記データストライプの2以上のセットに分けることができる、前記ディスクアレイと、前記上位装置及び前記ディスクアレイの間で転送されるデータの一時的記憶に用いられるキャッシュメモリと、前記ディスクアレイとのデータ転送を制御する下位側転送制御部とを備える。そして、前記ディスクアレイ制御部が、前記データディスクドライブ毎の故障発生の可能性を推定する推定部と、前記複数のデータディスクドライブの内から、前記推定された故障発生の可能性に応じて2以上のデータディスクドライブを、分割データコピーの対象として選び、前記選ばれた2以上のデータディスクドライブの各々から一つずつ分割記憶領域を分割することにより2以上の分割記憶領域を選び、前記選ばれた2以上の分割記憶領域は、前記パリティグループ内の前記データストライプの異なるセットにそれぞれ属し、そして、前記選ばれた2以上の分割記憶領域のデータを前記1以上のスペアディスクドライブへコピーするように前記下位側転送制御部及び前記キャッシュメモリを制御する分割データコピー部とを有する。

【0008】

一つの実施形態においては、前記ディスクアレイ制御部が、前記複数のデータディスクドライブの中から、前記推定された故障発生の可能性に応じて1つのデータディスクドライブを、ダイナミックスペアリングの対象として選び、前記選ばれた1つのデータディスクドライブから、前記分割データコピー部によるコピーが未だ行われていない残りの分割記憶領域を選び、そして、前記選ばれた残りの分割記憶領域のデータを前記スペアディスクドライブへコピーするように前記下位側転送制御部及び前記キャッシュメモリを制御するダイナミックスペアリング部を更に有する。

【0009】

一つの実施形態においては、前記分割データコピー部は、前記複数のデータディスクドライブの内の第1のデータディスクドライブの前記推定された故障発生の可能性が第1のレベルに達した場合、少なくとも、前記第1のデータディスクドライブと他の第2のデータディスクドライブとを、前記分割データコピーの対象として選択する。前記ダイナミックスペアリング部は、前記第1のデータディスクドライブの前記推定された故障発生の可能性が前記第1のレベルより高い第2のレベルに達した場合、前記第1のデータディスクドライブを、前記ダイナミックスペアリングの対象として選択する。

【0010】

一つの実施形態においては、前記分割データコピー部は、前記複数のデータディスクドライブの内の第1のデータディスクドライブの前記推定された故障発生の可能性が所定のレベルに達した場合、前記第1のデータディスクドライブと、前記推定された故障発生の可能性が前記第1のデータディスクドライブに次いで高い第2のデータディスクドライブとを、前記分割データコピーの対象として選択する。

【0011】

一つの実施形態においては、前記分割データコピー部は、前記選ばれた2以上の分割記憶領域のデータをコピーする過程で、前記選ばれた2以上の分割記憶領域から同時にデータを読み出すように前記下位側転送制御部及び前記キャッシュメモリを制御する。

【0012】

一つの実施形態においては、前記ディスクアレイ制御部が、前記選ばれた2以上の分割記憶領域のデータの前記1以上のスペアディスクドライブへのコピーが開始された後、前記上位装置からの前記選ばれた2以上の分割記憶領域への新たなデータの書き込み要求を前記上位側データ転送部から受けた場合、前記上位装置から受けた前記新たなデータを、前記選ばれた2以上の分割記憶領域へ書き込むと共に、前記スペアディスクドライブにも書き込むように前記下位側転送制御部及び前記キャッシュメモリを制御するスペアデータ最新化部を更に有する。

【0013】

一つの実施形態においては、前記分割データコピー部が、前記複数のデータディスクドライブから前記分割データコピーの対象として第1及び第2のデータディスクドライブを

選び、前記第1のデータディスクドライブから前記パリティグループの内の前側のデータストライプのセットに属する第1の分割記憶領域を選び、前記第2のデータディスクドライブから前記前側のデータストライプのセットに引き続く後側のデータストライプのセットに属する第2の分割記憶領域を選び、そして、前記第1と第2の分割記憶領域のデータを前記スペアディスクドライブへコピーするように前記下位側転送制御部及び前記キャッシュメモリを制御する。

【0014】

一つの実施形態においては、前記2以上のデータディスクドライブから選ばれる前記2以上の分割記憶領域が、実質的に等しいサイズをもつ。

【0015】

一つの実施形態においては、前記2以上のデータディスクドライブから選ばれる前記2以上の分割記憶領域が、前記2以上のデータディスクドライブの前記推定された故障発生の可能性に応じて異なるサイズを有する。

【0016】

一つの実施形態においては、前記推定部が、前記データディスクドライブ毎のエラー発生履歴を記憶し、前記記憶されたエラー発生履歴に基づいて前記データディスクドライブ毎に前記故障発生の可能性を推定する。

【0017】

本発明の別の観点に従うデータスペアリング方法は、上位装置と通信可能に接続され得るディスクアレイ装置であって、このディスクアレイ装置の全体の制御を行うディスクアレイ制御部と、前記上位装置とのデータ転送を制御する上位側転送制御部と、少なくとも、1つのパリティグループを構成する複数のデータディスクドライブと、1以上のスペアディスクドライブとを有するディスクアレイ装置であって、前記1つのパリティグループは前記複数のデータディスクドライブの記憶領域に渡って形成される多数のデータストライプを有し、前記多数のデータストライプは前記データストライプの2以上のセットに分けることができる、前記ディスクアレイと、前記上位装置及び前記ディスクアレイの間で転送されるデータの一時記憶に用いられるキャッシュメモリと、前記ディスクアレイとのデータ転送を制御する下位側転送制御部とを備えたものに適用される。この方法は、前記データディスクドライブ毎の故障発生の可能性を推定するステップと、前記複数のデータディスクドライブの内から、前記推定された故障発生の可能性に応じて、2以上のデータディスクドライブを、分割データコピーの対象として選ぶステップと、前記選ばれた2以上のデータディスクドライブの各々から一つずつ分割記憶領域を分割することにより2以上の分割記憶領域を選ぶステップであって、前記選ばれた2以上の分割記憶領域が、前記パリティグループ内の異なるデータストライプのセットにそれぞれ属するようになったステップと、前記選ばれた2以上の分割記憶領域のデータを前記1以上のスペアディスクドライブへコピーするように前記下位側転送制御部及び前記キャッシュメモリを制御することで分割データコピーを行うステップとを有する。

【0018】

この方法の一つの実施形態は、前記複数のデータディスクドライブの中から、前記推定された故障発生の可能性に応じて1つのデータディスクドライブを、ダイナミックスペアリングの対象として選ぶステップと、前記選ばれた1つのデータディスクドライブから、前記スペアディスクへのコピーが未だ行われていない残りの分割記憶領域を選び、前記選ばれた残りの分割記憶領域のデータを前記スペアディスクドライブへコピーするように前記下位側転送制御部及び前記キャッシュメモリを制御することでダイナミックスペアリングを行うステップを更に有する。

【0019】

一つの実施形態においては、前記2以上のデータディスクドライブを選ぶステップは、前記複数のデータディスクドライブの内の第1のデータディスクドライブの前記推定された故障発生の可能性が第1のレベルに達した場合に行なわれて、そこで、少なくとも前記第1のデータディスクドライブと他の第2のデータディスクドライブとが、前記分割デー

タコピーの対象として選択される。また、前記1つのデータディスクドライブを選ぶステップは、前記第1のデータディスクドライブの前記推定された故障発生の可能性が前記第1のレベルより高い第2のレベルに達した場合に行なわれて、そこで、前記第1のデータディスクドライブが、前記ダイナミックスペアリングの対象として選択される。

【0020】

本発明のまた別の観点に従う、データスペアリング制御装置は、上位装置と通信可能に接続され得るディスクアレイ装置であって、このディスクアレイ装置の全体の制御を行うディスクアレイ制御部と、前記上位装置とのデータ転送を制御する上位側転送制御部と、少なくとも、1つのパリティグループを構成する複数のデータディスクドライブと、1以上のスペアディスクドライブとを有するディスクアレイであって、前記1つのパリティグループは前記複数のデータディスクドライブの記憶領域に渡って形成される多数のデータストライプを有し、前記多数のデータストライプは前記データストライプの2以上のセットに分けることができる、前記ディスクアレイと、前記上位装置及び前記ディスクアレイの間で転送されるデータの一時記憶に用いられるキャッシュメモリと、前記ディスクアレイとのデータ転送を制御する下位側転送制御部とを備えたものに適用される。このデータスペアリング制御装置は、前記データディスクドライブ毎の故障発生の可能性を推定する推定部と、前記複数のデータディスクドライブの内から、前記推定された故障発生の可能性に応じて2以上のデータディスクドライブを、分割データコピーの対象として選び、前記選ばれた2以上のデータディスクドライブの各々から一つずつ分割記憶領域を分割することにより2以上の分割記憶領域を選ぶものであって、前記選ばれた2以上の分割記憶領域が前記パリティグループ内の異なるデータストライプのセットにそれぞれ属するようにする分割領域選択部と、前記選ばれた2以上の分割記憶領域のデータを前記スペアディスクドライブへコピーするように前記下位側転送制御部及び前記キャッシュメモリを制御する分割データコピー部とを備える。

【発明の効果】

【0021】

本発明に従うディスクアレイ装置によれば、或るパリティグループを形成する複数のデータディスクドライブの中から、それぞれの故障発生の可能性に応じて2以上のデータディスクドライブが選ばれ、それらのデータディスクドライブの各々の記憶領域が分割されて、そこから、異なるデータストライプのセットにそれぞれ属する2以上の分割記憶領域が選ばれ、それら2以上の分割記憶領域のデータが、それらのデータディスクドライブが故障する前に、スペアディスクドライブにコピーされる。このような分割データコピー処理が故障発生のある程度に高くなったデータディスクドライブについて行われることより、それらのデータディスクドライブの故障可能性が後に上昇した場合、分割データコピー処理が未だ行われていない残りの分割記憶領域についてのみ、例えばダイナミックスペアリング処理のようなデータ保全のための処理を行えばよいことになる。その結果、データ保全に要する処理時間が短縮できる。また、分割データコピー処理においては、2以上のディスクドライブの分割記憶領域から並列的にデータをREADするようにすれば、ディスクドライブのREAD時間も短縮される。結果として、高負荷環境においても多重ディスク故障によるデータロストの危険性が低減する。

【発明を実施するための最良の形態】

【0022】

以下、本発明に従うディスクアレイ装置の一実施形態を説明する。

【0023】

図1は、本発明の一実施形態に係るディスクアレイ装置を使用したコンピュータシステムの概略構成、及びこのディスクアレイ装置の内部の構成を示す。

【0024】

図1に示すように、ディスクアレイ装置100は、ホストコンピュータのような上位装置110と接続されており、複数のハードディスクドライブ（ドライブ番号#0～#N）300～330と1又は2以上のコントローラ200、210を備えている。また、この

ディスクアレイ装置 100 は、その操作・設定を行うための制御コンソール 120 を備え、コントローラ 200、210 が LAN インタフェース 201、211 を介して制御コンソール 120 にそれぞれ接続されている。

【0025】

コントローラ 200、210 はそれぞれ、上位装置 100 とのデータ転送を制御する上位側転送制御部 202、212、ディスクアレイ装置 100 全体の制御を行うディスクアレイ制御部 203、213、上位装置 100 及びディスクドライブ 300～330 のデータ転送等に用いられるキャッシュメモリ 230、及び、ディスクドライブ 300～330 のデータ転送を制御する下位側転送制御部 204、214 を備える。コントローラ 200 は、ポート 400～430 をそれぞれ介してディスクドライブ 300～330 にアクセスする。コントローラ 210 は、ポート 440～470 をそれぞれ介してディスクドライブ 300～330 にアクセスする。本実施形態の以下の説明では、複数のディスクドライブ 300～330 の物理的な並びは「ディスクアレイ」と呼ばれ、RAID 技術による冗長化された構成の論理ボリュームを「LU（論理ユニット）」と呼ばれる。

【0026】

次に、図 2 及び図 3 を用いて、本発明のディスクアレイ装置 100 が用いるディスクアレイの状態を管理するためのデータ構造について説明する。図 2 は、ディスクドライブ資源情報テーブル 500 を例示する模式図である。図 3 はスペアディスクドライブ資源情報テーブル 600 を例示する模式図である。

【0027】

図 2 に示されたディスクドライブ資源情報テーブル 500 は、コントローラ 200、210 のディスクアレイ制御部 203、213 の各々に存在する。ディスクドライブ資源情報テーブル 500 は、全てのディスクドライブ 300～330 の使用状況や容量などを管理するためのテーブルであり、全てのディスクドライブ 300～330 にそれぞれ対応する複数のディスクドライブ管理領域 510 を有する。各ディスクドライブ管理領域 510 に割り当てられた ROW（行）番号（ROW#0、ROW#1、…）及び COL（列）番号（COL#0、COL#1、…）は、対応するディスクドライブのディスクアレイ装置 100 内での位置を示している。

【0028】

各ディスクドライブ管理領域 510 は、対応するディスクドライブについて、例えば、使用状態 520、全容量 530、スペアリング済み容量（又はアドレス）540、及びエラーカウンタ 550 などの情報を有する。使用状態 520 には、対応するディスクドライブについて実装状況及び使用状況を示す値、例えば、未実装、実装（データディスクドライブとして使用されるか、スペアディスクドライブとして使用されるかも区別する）、及び未使用の 4 つの値が選択的に入る。全容量 530 は、対応するディスクドライブが実装されている場合に、そのディスクドライブの全記憶容量を表す値が入る。スペアリング済み容量（又はアドレス）540 は、対応するディスクドライブの故障発生可能性が高まった場合にスペアディスクドライブへのデータコピー時に用いられるフィールドであり、そのディスクドライブの全容量の中で現時点でスペアディスクドライブへのデータコピーが行われた容量（又はデータコピーが行われた記憶領域のアドレス（例えば、論理ブロック番号など））を示す値が入る。エラーカウンタ 550 には、対応するディスクドライブにおけるアクセスエラーの発生回数を示す値が入り、この値はディスクドライブでアクセスエラーが発生する都度にカウントアップ（又はカウントダウン）される。エラーカウンタ 550 の値は、後述するエラーレート（エラーの発生頻度）、換言すれば故障発生の可能性の高さを表すものとして用いられる。

【0029】

図 3 に示されたスペアディスクドライブ資源情報テーブル 600 は、コントローラ 200、210 のディスクアレイ制御部 203、213 の各々に存在する。スペアディスクドライブ資源情報テーブル 600 は、ディスクドライブ 300～330 の内、スペアディスクドライブとして使用されるものの状態を管理するためのテーブルであり、全てのスペアデ

ィスクドライブにそれぞれ対応する複数のスペアディスクドライブ情報領域 610 を有する。

【0030】

各スペアディスクドライブ情報 610 は、対応するスペアディスクドライブについての詳細情報、例えば、使用状態 620、スペア ROW 630、スペア COL 640、使用領域情報 [0] 650 ~ [1] 660 などの情報を含む。使用状態 620 は、対応するスペアディスクドライブの使用状態を示す値、例えば、未使用、一部使用、及び全使用の 3 つの値を選択的に持つ。使用状態 620 は、或るデータディスクドライブのデータのスペアリングを行う場合に、そのスペアディスクドライブの使用可否判断を行うために使用される。スペア ROW 630 及びスペア COL 640 には、対応するスペアディスクドライブのディスクアレイ装置 100 内での位置を示す ROW 番号及び COL 番号が入る。使用領域情報 [0] 650 ~ [m] 670 は、対応するスペアディスクドライブに含まれる m 個の記憶領域 ([0] ~ [m]) (m は、そのスペアディスクドライブの分割数であり、後述する図 4 に示す例では m = 2 である) にそれぞれ対応して設けられ、それら m 個の記憶領域がそれぞれのディスクドライブのデータ復元用に使用されたかを管理するための情報である。各使用領域情報は、使用中フラグ 670、開始アドレス 680、終了アドレス 690、コピー元 ROW 700、及びコピー元 COL 710 などの情報を有する。使用中フラグ 670 は、対応する記憶領域が現在使用中であるか否かを示す。開始アドレス 680 及び終了アドレス 690 は、対応する記憶領域の開始アドレスと終了アドレスを示す。コピー元 ROW 700 及びコピー元 COL 710 には、対応する記憶領域にコピーされたデータの元々記憶されてたデータディスクドライブのディスクアレイ装置 100 内での位置を示す ROW 番号と COL 番号が入る。

【0031】

ディスクアレイ装置 100 のディスクアレイ制御部 203、213 (図 1 参照) は、上述したディスクドライブ資源情報テーブル 500 及びスペアディスクドライブ資源情報テーブル 600 を用いて、ディスクドライブ 300 ~ 330 の状態を把握し、そして、ディスクドライブの故障からデータを保護するために、以下に説明する制御動作を行う。

【0032】

以下の説明では、一例として、各ディスクドライブのエラーレート (アクセスエラーの発生頻度) が、各ディスクドライブの故障発生の可能性を示す 3 段階のレベルに分類される。すなわち、「レベル 1」は故障発生確率が低いことを、「レベル 2」は故障発生確率が中程度であることを、そして、「レベル 3」は故障発生確率が高いことを意味する。ディスクアレイ制御部 203、213 は、或るディスクドライブのエラーレートが例えばレベル 2 になったとき、そのディスクドライブについて本発明の原理に従う「スペアディスクへの分割データコピー」を行なって、故障発生に備える。その後、そのディスクドライブでさらに多くのエラーが発生して、エラーレートが例えばレベル 3 になると、ディスクアレイ制御部 203、213 は、そのディスクドライブについてダイナミックスペアリングを実施し、そして、ダイナミックスペアリングの完了後、そのディスクドライブを閉塞する。

【0033】

このようなディスクドライブの故障からデータを保護するための制御について、以下により具体的に説明する。以下の説明では、RAID-5 技術に従い 3D+1P 形式のパリティグループを形成するように構成されたディスクアレイの場合を例にとるが、RAID 技術に従う他の構成のディスクアレイにおいても、本発明の原理に従う制御が適用できることは当業者は容易に理解できるはずである。

【0034】

図 4 は、3D+1P 形式のパリティグループを形成するディスクアレイの構成例を示す。

【0035】

図 4 に示す例では、3D+1P 形式の 1 つのパリティグループを形成する 4 つのデータディスクドライブ (DDD#0 ~ DDD#3) 800 ~ 830 からなるディスクアレイが存在する。この

ディスクアレイ 800～830 内に、例えば 2 つの論理ユニット LU0、LU1 が定義されている。一方の論理ユニット LU0 は論理ブロック Block0～Block17 から構成され、他方の論理ユニット LU1 は論理ブロック Block18～BlockX で構成されている（1 論理ブロックのサイズは例えば 64 K バイトである）。これら 4 つの異なるデータディスクドライブ 800～830 にそれぞれ含まれる図中横方向一列に並ぶ 4 つの論理ブロックが、3D+1P 形式の一つのデータストライプを形成する。ここで、各データストライプ内の P(Q-R) と記された論理ブロックは、同じデータストライプ内の他の論理ブロック BlockQ～BlockR のデータのパリティデータを記憶したブロックであることを意味する。例えば、第 1 のデータストライプは、3 つのデータ論理ブロック Block0～Block2 と 1 つのパリティ論理ブロック P(0-2) から構成され、また、第 2 のデータストライプは、3 つのデータブロック Block3～Block5 と 1 つのパリティブロック P(3-5) から構成される。このようにして、1 つのパリティグループを形成するディスクアレイ 800～830 内には、多数のデータストライプが存在する。さらに、このディスクアレイ 800～830 の他に、少なくとも 1 つのスペアディスクドライブ (SDD#A) が存在する。

【0036】

図 5 は、図 4 に例示したような構成のディスクアレイ 800～830 に適用される、ディスクアレイ制御部 203、213（図 1 参照）による、データディスクドライブの故障からデータを保護するための制御の全体的な流れを示す。

【0037】

図 5 に示すように、ディスクアレイ制御部 203、213 は、ディスクアレイ 800～830 内に、そのエラーレートがレベル 2 に達したディスクドライブが存在する否かチェックする（ステップ 900）。このチェックは、図 2 に示されたディスクドライブ資源情報テーブル 500 内から、そのエラーカウンタ 550 の値がレベル 2 に相当する所定の閾値に達したディスクドライブ管理領域 510 を探すことで行うことができる。そのようなディスクドライブ管理領域 510 が見つかった場合、そのディスクドライブ管理領域 510 に対応する ROW 番号と COL 番号により決まる位置にあるディスクドライブが、エラーレートがレベル 2 に達したディスクドライブとして検出される。そのようなディスクドライブが検出された場合、ディスクアレイ制御部 203、213 は、図 3 に示されたスペアディスクドライブ資源情報テーブル 600 を参照し、使用可能なスペアディスクドライブの有無を判定する（ステップ 910）。使用可能なスペアディスクが有る場合、ディスクアレイ制御部 203、213 は、下位側転送制御部 204、214 及びキャッシュメモリ 220 を制御して、上記検出されたディスクドライブについて、分割データコピー処理を行う（ステップ 920）。分割データコピー処理の具体的方法については後に説明する。他方、使用可能なスペアディスクがない場合には、ディスクアレイ制御部 203、213 は、制御コンソール 110 に対して、ディスクドライブのエラーレート上昇を知らせる警告を発し、オペレータの注意を促す（ステップ 950）。

【0038】

また、ディスクアレイ制御部 203、213 は、ディスクアレイ 800～830 内に、そのエラーレートがレベル 3 に達したディスクドライブが存在する否かチェックする（ステップ 930）。このチェックは、図 2 に示されたディスクドライブ資源情報テーブル 500 内から、そのエラーカウンタ 550 の値がレベル 3 に相当する所定の閾値に達したディスクドライブ管理領域 510 を探すことで行うことができる。そのようなディスクドライブ管理領域 510 が見つかった場合、そのディスクドライブ管理領域 510 に対応する ROW 番号と COL 番号により決まる位置にあるディスクドライブが、エラーレートがレベル 3 に達したディスクドライブとして検出される。そのようなディスクドライブが検出された場合、ディスクアレイ制御部 203、213 は、下位側転送制御部 204、214 及びキャッシュメモリ 220 を制御して、その検出されたディスクドライブについてダイナミックスペアリングを実行する（ステップ 940）。ここで、上述の説明から明らかなように、ダイナミックスペアリングの対象となったディスクドライブについては、通常、既にステップ 920 の分割データコピーが行われている。後述する分割データコピーの具体的な

処理から分かるように、分割データコピーによって、そのディスクドライブ内の全記憶領域の内の半分の記憶領域のデータがスペアディスクドライブに既にコピーされ終わっている。そのため、ダイナミックスペアリングでは、残り半分の記憶領域のデータをスペアディスクドライブへコピーするだけでよい。よって、ダイナミックスペアリングにかかる時間は、従来よりも半減する。ダイナミックスペアリングの完了後、そのディスクドライブは閉塞され、正常なディスクドライブへの交換作業が行われる。図5には示されていないが、閉塞されたディスクドライブの交換完了後、そのディスクドライブにスペアディスクドライブからデータのコピーバック（データ復元）が行われる。このデータ復元が完了した後、そのスペアディスクドライブは待機状態に設定され、次の分割コピー処理及びダイナミックスペアリングの機会に利用される。

【0039】

次に、図5中のステップ920の分割データコピー処理の具体的方法について説明する。

【0040】

図4には、既に説明したディスクアレイ800～830の構成例と共に、このディスクアレイ800～830において実行される分割データコピー処理の具体的な態様の一例が示されている。図4を参照して、この分割データコピー処理の一例を説明する。

【0041】

図4に示されたディスクアレイ800～830において、例えば、一つのデータディスクドライブ（DDD#0）800のエラーレートがレベル2に達したとする。このとき、他のデータディスクドライブ（DDD#1～#3）810～830のエラーレートはレベル1であったとする。ただし、レベル1のデータディスクドライブ（DDD#1～#3）810～830の中で、データディスクドライブ（DDD#1）810のエラーカウンタの値が最も高かったとする。また、このとき、使用可能なスペアディスク（SDD#A）840が存在していたとする。

【0042】

このような状況では、レベル2に達したディスクドライブ（DDD#0）800について、それが故障してREAD動作が不可能になる前に、分割データコピー処理が実行される。この分割データコピー処理では、ディスクドライブ（DDD#0）800の全記憶領域の中から、図中上半分（前半）の記憶領域（論理ブロックBlock0～Block12）850が選ばれ、そして、選ばれた上半分記憶領域850のデータを、スペアディスクドライブ（SDD#A）840の対応する上半分記憶領域880へコピーする処理が実行される。すなわち、下位側転送制御部204又は214の制御により、上半分記憶領域850のデータがキャッシュメモリ220に読み出され、そして、キャッシュメモリ220からスペアディスクドライブ（SDD#A）840の対応する上半分記憶領域880へ書き込まれる。これと並行して、このディスクドライブ（DDD#0）800に次いで高いエラーカウント値をもつ（つまり、次に故障発生の可能性の高い）別の一つのディスクドライブ（DDD#1）810が選ばれ、選ばれたディスクドライブ（DDD#1）810の中から、上述したディスクドライブ（DDD#0）800内の分割コピー範囲（つまり、上半分記憶領域850）が属する上半分（前半）のデータストライプのセットに引き続く下半分（後半）のデータストライプのセットに該当する下半分（後半）の記録場所領域、（Block16、P7、Block21、…）870が選ばれ、そして、この選ばれた下半分記録場所領域（Block16、P7、Block21、…）870のデータを、スペアディスクドライブ（SDD#A）840の対応する半分記憶領域890へコピーする処理が実行される。すなわち、下位側転送制御部204又は214の制御により、下半分記録場所領域870のデータがキャッシュメモリ220に読み出され、そして、キャッシュメモリ220からスペアディスクドライブ（SDD#A）840の対応する半分記憶領域890へ書き込まれる。

【0043】

要するに、分割データコピー処理では、1つのパリティグループを形成する複数のディスクアレイ800～830の中から、ある程度高い故障発生可能性をもつ2つのディスク

ドライブ (DDD #0、#1) 8 0 0、8 1 0 が選ばれ、選ばれた 2 つのディスクドライブ (DD #0、#1) 8 0 0、8 1 0 から、異なるデータストライプのセットにそれぞれ属する上半分 (前半) と下半分 (後半) の記憶領域 8 5 0、8 7 0 がそれぞれ選ばれ、そして、選ばれた上半分 (前半) と下半分 (後半) の記憶領域 8 5 0、8 7 0 から同時にデータが READ されて、スペアディスク 8 4 0 に Write される。このとき、同時に 2 つのディスクドライブ (DDD#0、#1) 8 0 0、8 1 0 から READ されたデータは、ディスクアレイ制御部 2 0 3、2 1 3 によってスケジューリングされ、スペアディスクドライブ 8 4 0 へ WRITE される。この 2 つのディスクドライブからの同時データ READ によって、データ READ にかかる時間が半減する。これによるコピー時間の短縮は特に高負荷環境においては有利である。さらに、スペアディスクドライブ 8 4 0 が複数のアクセスポートを持つ場合は、スペアディスクドライブ 8 4 0 への多重データ WRITE を行うことができるから、分割データコピー処理はより高速化される。

【0 0 4 4】

なお、分割データコピー処理の実行中、コピー対象のディスクドライブ (DDD#0、#1) 8 0 0、8 1 0 内の既にコピーが完了した論理ブロックに対して新たなデータの WRITE 要求が上位装置 1 1 0 などから上位側転送制御部 2 0 2、2 1 3 を通じて入力された場合には、ディスクアレイ制御部 2 0 3、2 1 3 は、下位側転送制御部 2 0 4 とキャッシュメモリ 2 2 0 を制御して、Write 対象のディスクドライブ (DDD#0 又は #1) 8 0 0 又は 8 1 0 とスペアディスクドライブ (SDD#A) 8 4 0 の双方へ当該データを 2 重に Write する。これにより、スペアディスクドライブ (SDD#A) 8 4 0 内のデータは常に最新状態に保たれ、コピー元のディスクドライブ (DDD#0、#1) 8 0 0、8 1 0 の故障発生に備えることができる。

【0 0 4 5】

図 6 は、分割データコピー処理の具体的な流れを示す。

【0 0 4 6】

図 6 に示すように、分割データコピー処理が開始されると、エラーレートがレベル 2 に達したディスクドライブと、このディスクドライブに次いでエラーレート (エラーカウンタ値) が高いディスクドライブから、分割データコピーの対象である 2 つの記憶領域が選択される (ステップ 9 6 0)。図 4 の例では、2 つのディスクドライブ (DDD#0 又は #1) 8 0 0 又は 8 1 0 から、半分記憶領域 8 5 0 と 8 6 0 がそれぞれ選ばれるのである。ここで、重要な点は、選ばれた 2 つの記憶領域 8 5 0 と 8 6 0 は、それが属するデータストライプのセットにおいて重複しないことである。換言すれば、コピーの対象として選ばれた全ての論理ブロックは、それぞれ異なるデータストライプのセットに属するというのである。これにより、出来る限り多くのデータが保護される可能性が高まるのである。

【0 0 4 7】

この後、上述した 2 つのディスクドライブについて並行して、それぞれのコピー対象領域内の全ての論理ブロックからデータが逐次に Read され、Read されたデータが所定のスケジュールに従ってスペアディスクドライブの対応する論理ブロックに Write される (一方のドライブについてはステップ 9 7 0 及びステップ 9 9 0 ~ 1 0 1 0 のループ、他方のドライブについてはステップ 9 7 0 及びステップ 1 0 2 0 ~ 1 0 5 0 のループ)。

【0 0 4 8】

図 7 は、エラーレートがレベル 3 に達したディスクドライブについて行われるダイナミックスペアリング処理の具体的な流れを示す。

【0 0 4 9】

図 7 に示すように、ダイナミックスペアリング処理が開始されると、この処理の対象となるディスクドライブの中で既にスペアリング (スペアディスクへのコピー) が完了している記憶領域がどこであるかが、ディスクアレイ制御部 2 0 3 又は 2 1 3 によって確認される (ステップ 1 1 0 0)。この確認は、図 2 に示したディスクドライブ資源情報テーブル 5 0 0 内のそのディスクドライブに対応するディスクドライブ管理領域 5 1 0 のスペアリング済容量 (又はアドレス) 5 4 0 を参照することで行うことができる。通常、そのデ

ィスクドライブについては分割データコピー処理が既に行われているから、そのディスクドライブの中の半分の記憶領域は既にスペアリングが完了していることになる。その後、ディスクアレイ制御部203又は213の制御により、そのディスクドライブの中の未だスペアリングが完了していない記憶領域のデータが、ブロック番号の順で逐次に、下位側転送制御部204又は214を通じてキャッシュメモリ220に読み出され、キャッシュメモリ220から下位側転送制御部204又は214を通じてスペアディスクドライブの対応する記憶領域にコピーされる（ステップ1110～1140のループ）。

【0050】

なお、ダイナミックスペアリングが完了する前に、ダイナミックスペアリングの対象のディスクドライブが故障してデータのREADができなくなった場合は、Read対象のデータと同じパリティグループを構成する他のディスクドライブ内のデータを用いたコレクションコピー動作により、そのRead対象のデータが復元されてスペアディスクドライブへコピーされることになる。このコレクションコピー動作が完了する前に、上述した他のディスクドライブも故障してデータのREADができなくなった場合は、従来技術によれば、これら故障したディスクドライブに関わるサブシステムはシステムダウン状態に陥らざるを得ない。しかしながら、本実施形態では、その故障した他のディスクドライブについて既に分割データコピーが行われていることが多いため、上記のような二重故障が発生しても、スペアディスクドライブ内のデータを用いて、コレクションコピーの対象のデータを復元することができる。

【0051】

次に、図8を参照して、3D+1P形式のディスクアレイを構成する4つのデータディスクドライブに対して2つのスペアディスクドライブが用意されている場合における、上述したデータ保護のための制御動作の具体例を説明する。

【0052】

図8Aに示すように、3D+1のディスクアレイを構成する4つのデータディスクドライブDD#0～#3において、ディスクドライブDDD#0のエラーレートがレベル2に達し、他のディスクドライブDDD#1～#3のエラーレートがレベル1（ただし、#2>#3>#1）であった場合、レベル2のエラーレートをもつディスクドライブDDD#0と、次に高いエラーレートをもつディスクドライブDDD#2について、スペアディスクドライブSDD#Aへの分割データコピー処理が実行される。すなわち、ディスクドライブDDD#0の上半分（前半）のデータ#0_UHとディスクドライブDDD#2の下半分（後半）のデータ#2_LHが、スペアディスクドライブSDD#Aへコピーされる。

【0053】

その後、図8Bに示すように、ディスクドライブDDD#0のエラーレートが上昇してレベル3に達した場合、このディスクドライブDDD#0がREADできなくなる前に、このディスクドライブDDD#0についてスペアディスクドライブSDD#Aへダイナミックスペアリング処理が実行される。このダイナミックスペアリング処理では、ディスクドライブDDD#0の未だスペアリングされてない下半分（後半）のデータ#0_LHをスペアディスクドライブSDD#Aにコピーするだけでよい。このとき、スペアディスクドライブSDD#A内にあったディスクドライブDDD#2の下半分（後半）のデータ#2_LHは消去されることになる。このとき、ディスクドライブDDD#2のエラーレートもレベル2に達した場合、分割データコピーが実行されて、ディスクドライブDDD#2の上半分（前半）のデータ#2_UHと次にエラーレートの高いディスクドライブDDD#3の下半分（後半）のデータ#3_LHが、もう一つのスペアディスクドライブSDD#Bへコピーされる。

【0054】

ディスクドライブDDD#0のダイナミックスペアリング処理の完了後、ディスクドライブDD#0は閉塞され、別の正常なディスクドライブに交換される。分割データコピー処理所路ディスクドライブDDD#2、DDD#3は、分割データコピー処理が完了した後もエラーレートがレベル3に達しない限り、通常通り使用される。ただし、ディスクドライブDDD#2又はDDD#3へのデータのWRITEが行われるときには、ディスクドライブDDD#2又はDDD#3と共にスペ

アディスクドライブSDD#Bへも同じデータが2重にWriteされ、それにより、スペアディスクドライブSDD#B内のデータが常に最新に保たれる。

【0055】

ディスクドライブDDD#0交換が完了した後、図8Cに示すように、スペアディスクドライブSDD#AからディスクドライブDDD#0へデータコピーが行われて、ディスクドライブDDD#0のデータが回復される。この回復処理が完了した後、スペアディスクドライブSDD#Aは待機状態となる。この段階で、スペアディスクドライブSDD#Bのデータは引き続き最新の状態に維持され、ディスクドライブDDD#2、DDD#3の将来のエラーレート上昇に備える。

【0056】

図9は、図8に示したデータコピー動作の変形例を示す。

【0057】

上述した図8のデータコピー動作では、図8Aに示された分割データコピー処理において、2つのディスクドライブDDD#0、DDD#2のデータ#0_UH、#2_LHが、同一のスペアディスクドライブSDD#Aへコピーされる。しかし、変形例として、図9Aに示すように、2つのディスクドライブDDD#0、DDD#2のデータ#0_UH、#2_LHをそれぞれ異なるスペアディスクドライブSDD#A、SDD#Bにコピーすることもできる。そのようにすると、その後に、図9Bに示すようにディスクドライブDDD#0のエラーレートがレベル3に達してダイナミックスペアリング処理を行う場合、既にスペアリングされたディスクドライブDDD#2のデータ#2_LHを消去することなしに、ディスクドライブDDD#0の残りデータ#0_LHをスペアディスクドライブSDD#Aにコピーすることができる。その結果、ディスクドライブDDD#2のエラーレートがレベル2に達して分割データコピー処理を行う場合、ディスクドライブDDD#2のスペアリング済みデータ#2_LHのコピーを省略して、次に高いエラーレートをもつディスクドライブDDD#3のデータ#3_UHだけをコピーすることができる。

【0058】

図10は、図8に示したデータコピー動作のまた別の変形例を示す。

【0059】

図10Aに示した分割データコピー処理の動作は、図8Aに示したものと同一である。しかし、図10Bに示したディスクドライブDDD#0のダイナミックスペアリング処理においては、ディスクドライブDDD#0の下半分（後半）のデータ#0_LHを、上半分（前半）のデータ#0_UHとは別のスペアディスクドライブSDD#Bにコピーする。これにより、既にスペアリングされたディスクドライブDDD#2のデータ#2_LHを消去しなくてよい。その結果、図9の動作例と同様に、ディスクドライブDDD#2のエラーレートがレベル2に達して分割データコピー処理を行う場合、ディスクドライブDDD#2のスペアリング済みデータ#2_LHのコピーを省略して、次に高いエラーレートをもつディスクドライブDDD#3のデータ#3_UHだけをコピーすることができる。

【0060】

図11は、図4に示した分割データコピー処理の動作の変形例を示す。

【0061】

図4に示した分割データコピー処理では、2つのディスクドライブDDD#0、DDD#1から実質的に均等なサイズの（つまり、半分サイズの）記憶領域805、807をそれぞれ選んで、それらのデータのスペアリングを行っている。変形例として、図11に示すように、2つのディスクドライブDDD#0、DDD#1から不均等なサイズの記憶領域1200、1210をそれぞれ選んで、それらのデータのスペアリングを行うこともできる。図11の例では、エラーレート（故障発生の可能性）のより高いディスクドライブDDD#0からより大きいサイズの記憶領域1200をコピーし、エラーレートのより低いディスクドライブDDD#1からより小さいサイズの記憶領域1210をコピーしている。このような不均等分割データコピー処理によれば、図4に示したような均等分割データコピー処理に比べて、分割データコピーに要する時間の短縮効果は劣るが、その分、後に行われるダイナミックスペアリング処理（エラーレートのより高いディスクドライブDDD#0について行われる可能性が高い）において処理時間が短縮できる効果が期待できる。

【0062】

図12は、図4に示した分割データコピー処理の動作の別の変形例を示す。

【0063】

図4に示した分割データコピー処理では、2つのディスクドライブDDD#0、DDD#1が処理対象として選ばれる。変形例として、図12に示すように、3つ又はそれより多くのディスクドライブDDD#0、DDD#1、DDD#2を分割データコピー処理対象として選んでもよい。図12の例では、エラーレートの高い順に3つのディスクドライブDDD#0、DDD#1、DDD#2が選ばれ、それらからパリティグループが重ならない3つの記憶領域1300、1310、1320がそれぞれ選ばれてコピーされる。

【0064】

以上、本発明の実施形態を説明したが、この実施形態は本発明の説明のための例示にすぎず、本発明の範囲をこの実施形態にのみ限定する趣旨ではない。本発明は、その要旨を逸脱することなく、その他の様々な態様でも実施することができる。

【0065】

例えば、上述した実施形態では、エラーレートを3段階のレベルに分類して、レベル2のドライブについて分割データコピーを行い、レベル3のドライブについてダイナミックスペアリングを行う。しかし、エラーレートを3段階より多くのレベルに分類して、中間的な複数のレベル（例えば、4段階のレベルのうちのレベル2と3）のドライブについて、レベルに応じた分割データコピーを行う（例えば、レベルが高いほど、スペアリングされるデータサイズが増える）ようにすることもできる。

【0066】

また、上述の実施形態では、「RAID-5」技術に従う3D+1P形式のパリティグループを構成するディスクアレイを例にとり説明したが、「RAID-1」～「RAID-5」のいずれかに従う他の形式のパリティグループを構成するディスクアレイにも本発明が適用可能なことは明らかである。例えば、3D+1P形式のパリティグループであっても、パリティデータの配置などにおいて、図13に示す例やその他のバリエーションが存在するが、それらバリエーションのいずれにも本発明は適用可能である。

【0067】

また、上記実施形態では、分割データコピー及びダイナミックスペアリングなどの動作を、ディスクアレイ装置が、上位装置などの外部装置からの指令によらずに自立的に制御しているが、外部装置からの指令によりそのような制御を行えるようにしてもよい。

【図面の簡単な説明】

【0068】

【図1】 本発明の一実施形態にかかるディスクアレイ装置を使用したコンピュータシステムの概略構成、及びこのディスクアレイ装置の内部の構成を示したブロック図。

【図2】 ディスクドライブ資源情報テーブルの構造を示す模式図。

【図3】 スペアディスクドライブ資源情報テーブルの構造を示す模式図。

【図4】 3D+1P形式のパリティグループを形成するディスクアレイの構成と、スペアディスクドライブへの分割データコピーの様子を示した模式図。

【図5】 ディスクアレイ制御部203、213が行うディスクアレイのデータ保護のための制御動作の流れを示すフローチャート。

【図6】 分割データコピー処理の具体的な流れを示すフローチャート。

【図7】 ダイナミックスペアリング処理の具体的な流れを示すフローチャート。

【図8】 3D+1P形式のディスクアレイを構成する4つのデータディスクドライブに対して2つのスペアディスクドライブが用意されている場合における、データ保護のための制御動作の具体例を示す模式図。

【図9】 図8に示したデータコピー動作の変形例を示す模式図。

【図10】 図8に示したデータコピー動作の別の変形例を示す模式図。

【図11】 分割データコピー処理の変形例を示す模式図。

【図12】 分割データコピー処理の別の変形例を示す模式図。

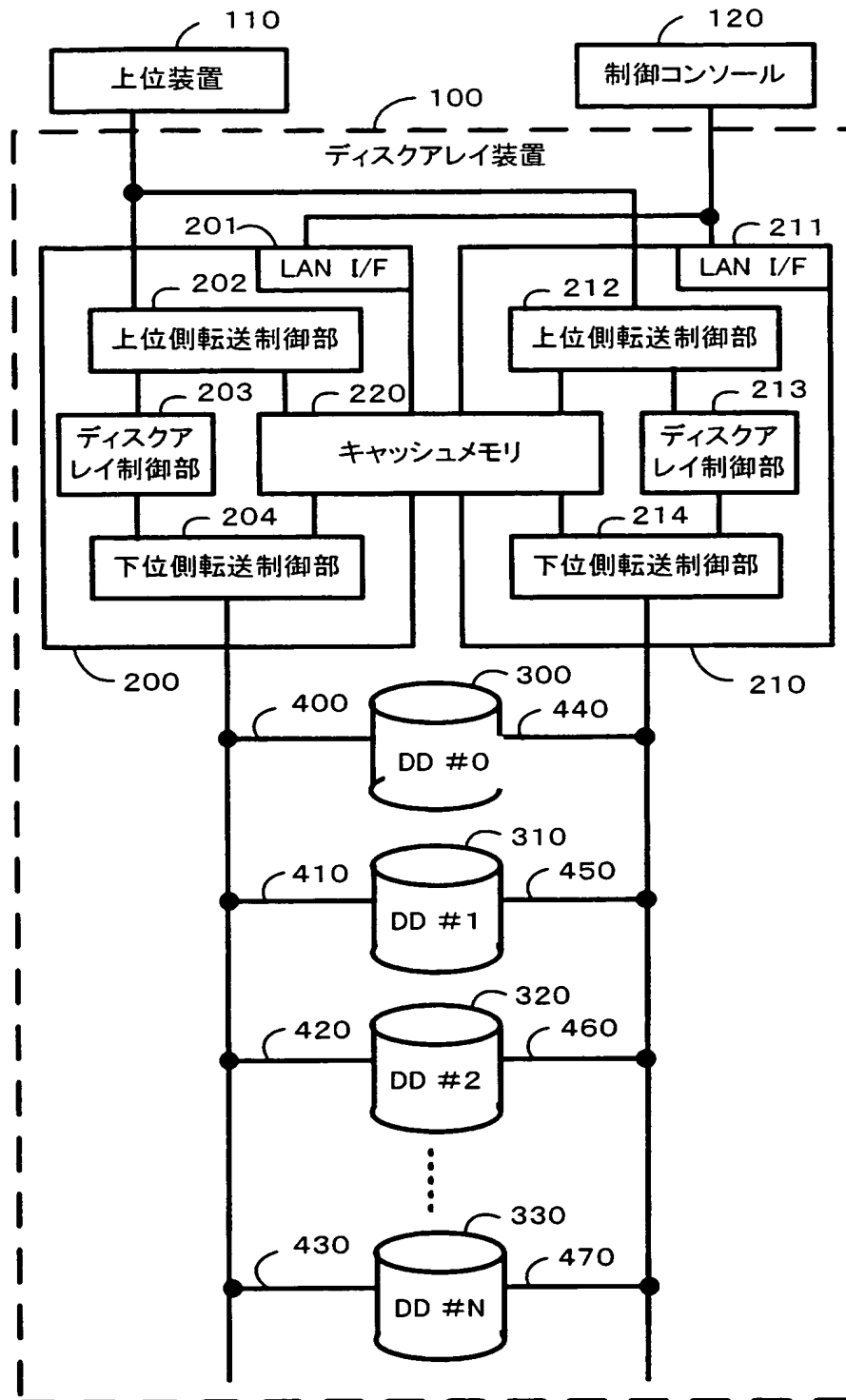
【図 13】 パリティグループの構成の変形例を示す模式図。

【符号の説明】

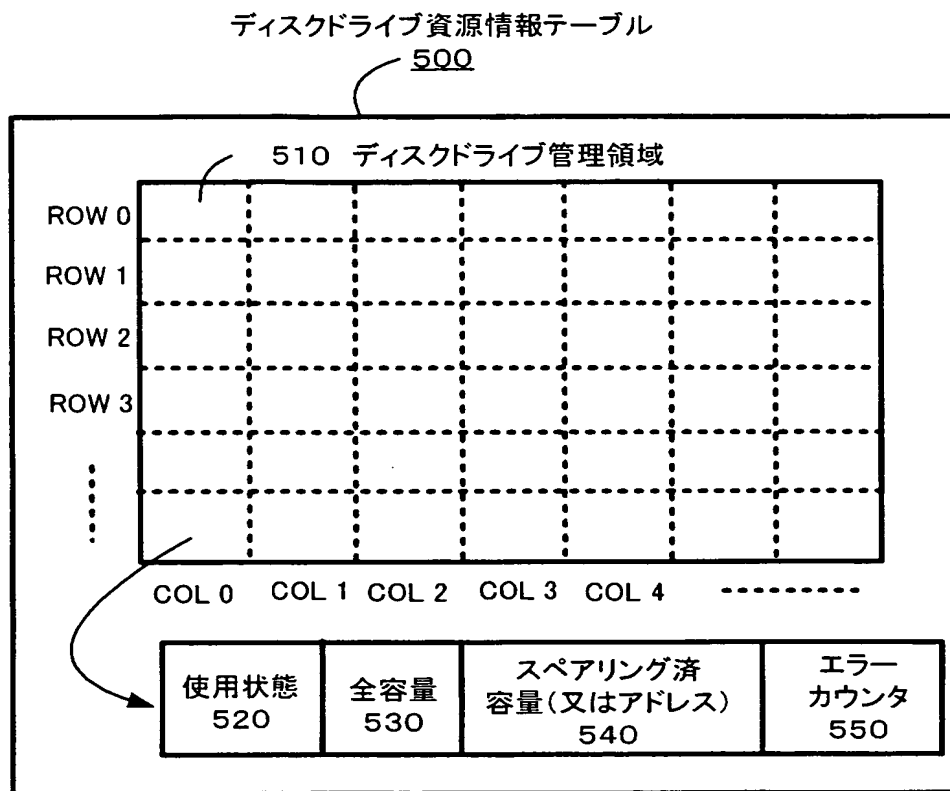
【0069】

100…ディスクアレイ装置
 110…上位装置
 120…制御コンソール
 200、210…コントローラ
 202、212…上位側転送装置
 203、213…ディスクアレイ制御部
 220…キャッシュメモリ
 204、214…下位側転送装置
 300～330…ディスクドライブ#0～#N
 500…ディスクドライブ資源情報テーブル
 510…ディスクドライブ管理領域
 520…使用状態
 530…全容量
 540…スオペアリング済容量（又はアドレス）
 550…エラーカウンタ
 600…スペアディスクドライブ資源情報テーブル
 610…スペアディスクドライブ情報領域
 620…使用状態
 630…スペアROW
 640…スペアCOL
 650～670…使用領域情報[0]～[m]
 670…使用中フラグ
 680…開始アドレス
 690…終了アドレス
 700…コピー元ROW
 710…コピー元COL
 800～830…データディスクドライブ#0～#3
 840…スペアディスクドライブ#A
 850、1200、1300…ディスクドライブ#0の分割コピー元の記憶領域
 860…ディスクドライブ#0の分割コピー対象外の記憶領域
 870、1210、1310…ディスクドライブ#1の分割コピー元の記憶領域
 880、890、1220、1230、1340～1350…スペアディスクドライブ#A
 の分割コピー先の記憶領域
 1320…ディスクドライブ#2の分割コピー元の記憶領域

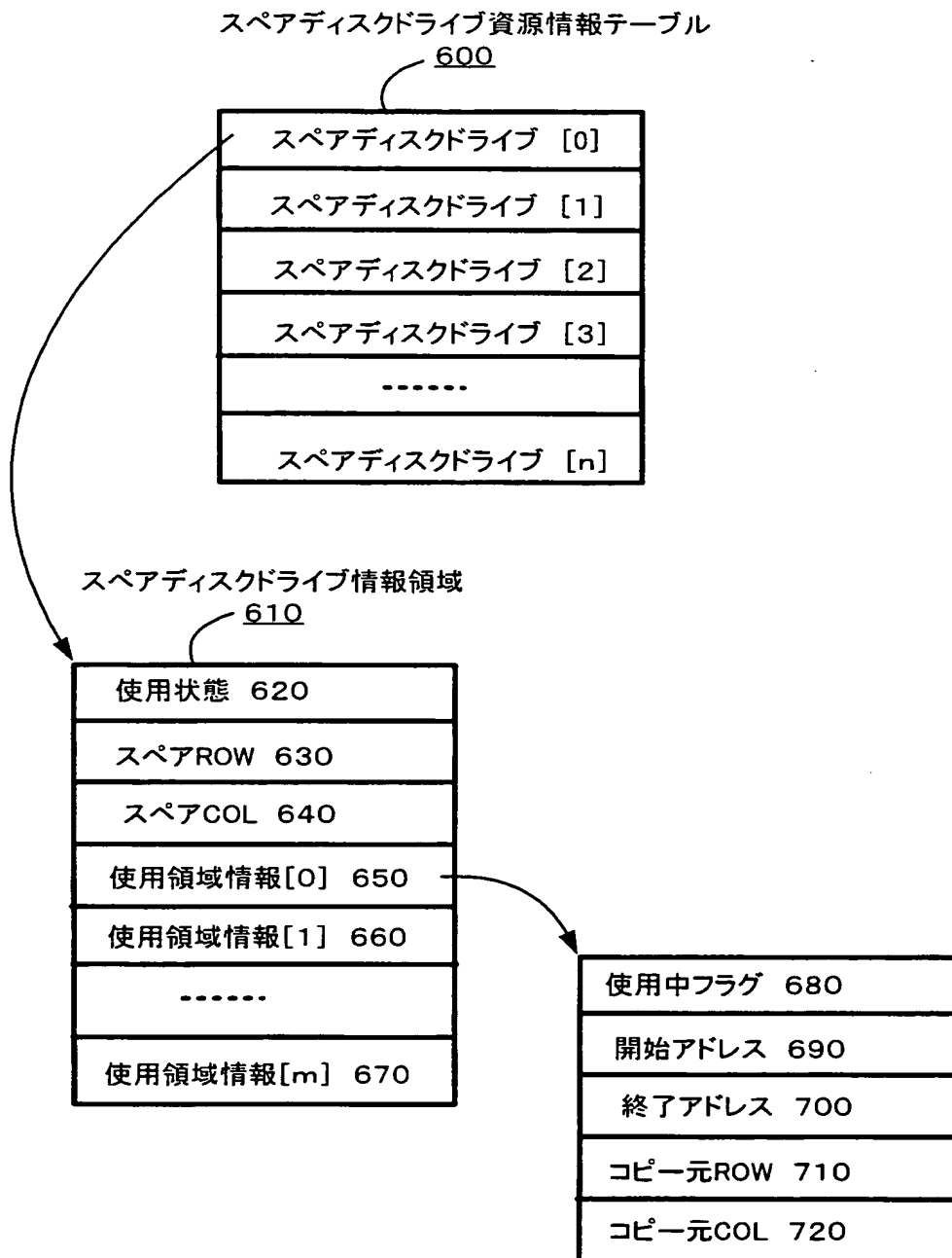
【書類名】図面
【図 1】



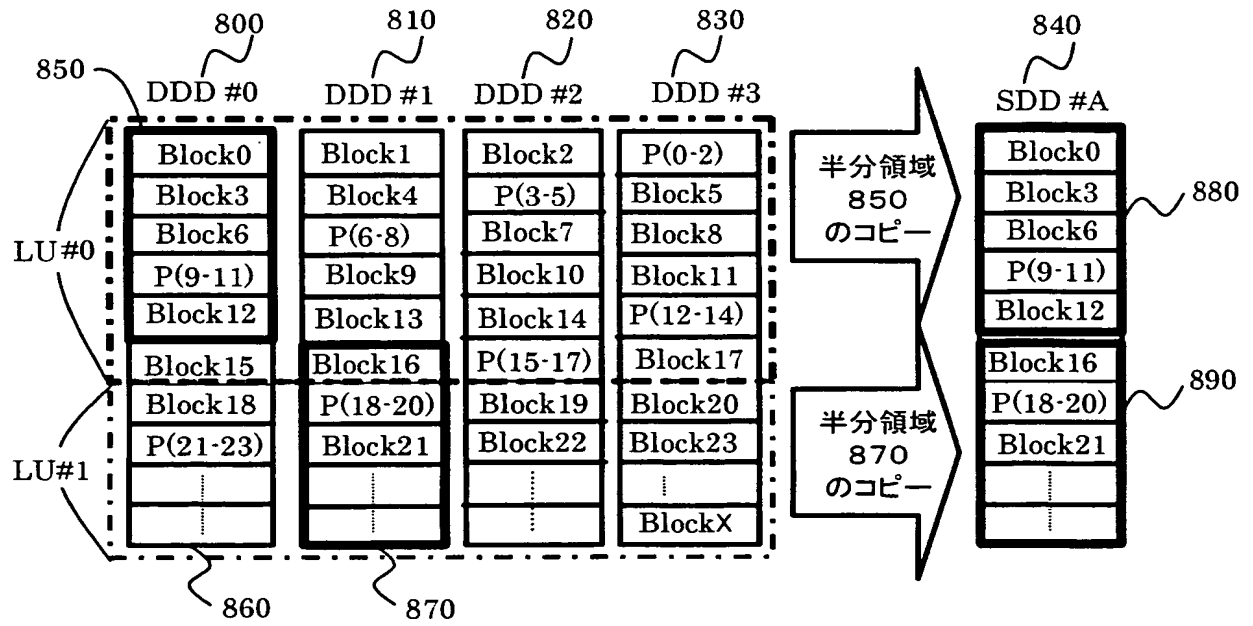
【図 2】



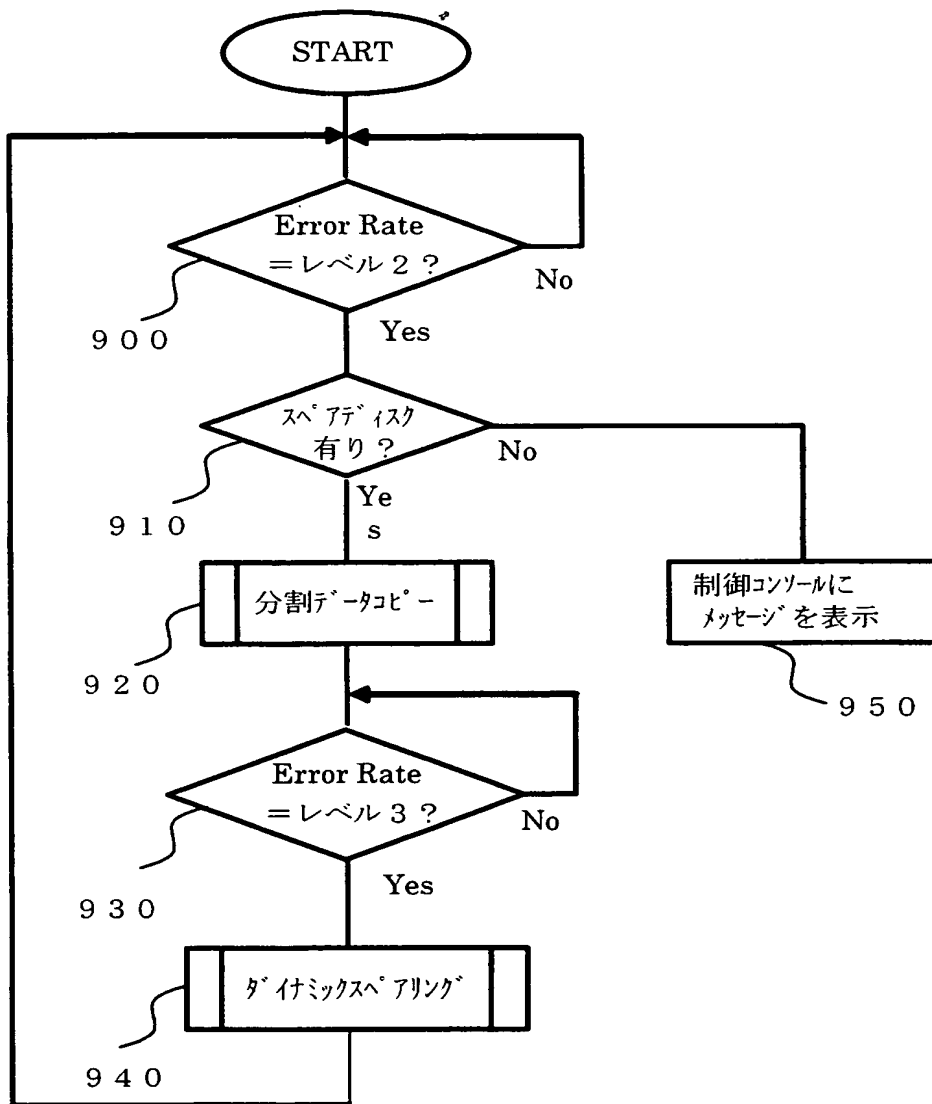
【図 3】



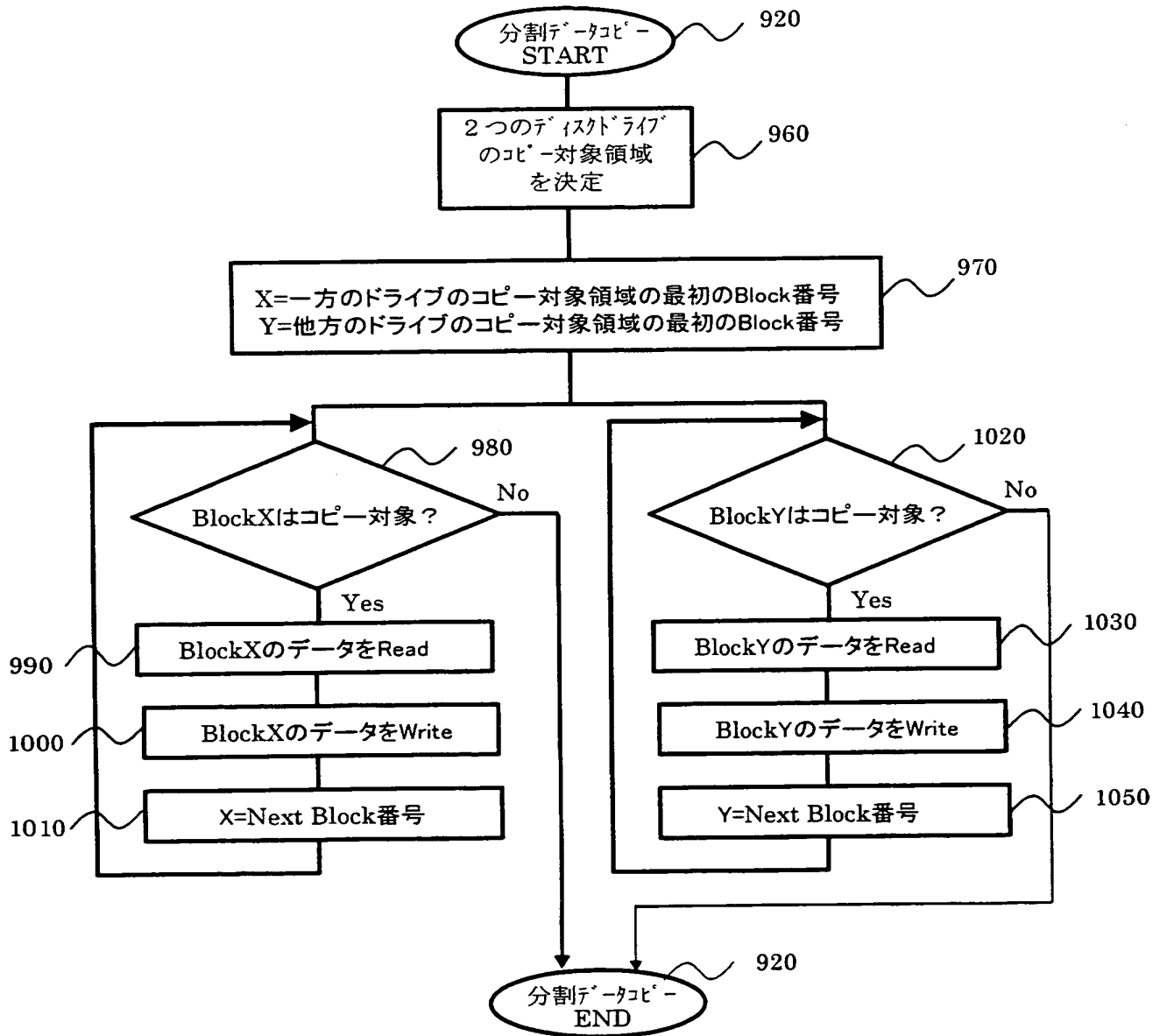
【図 4】



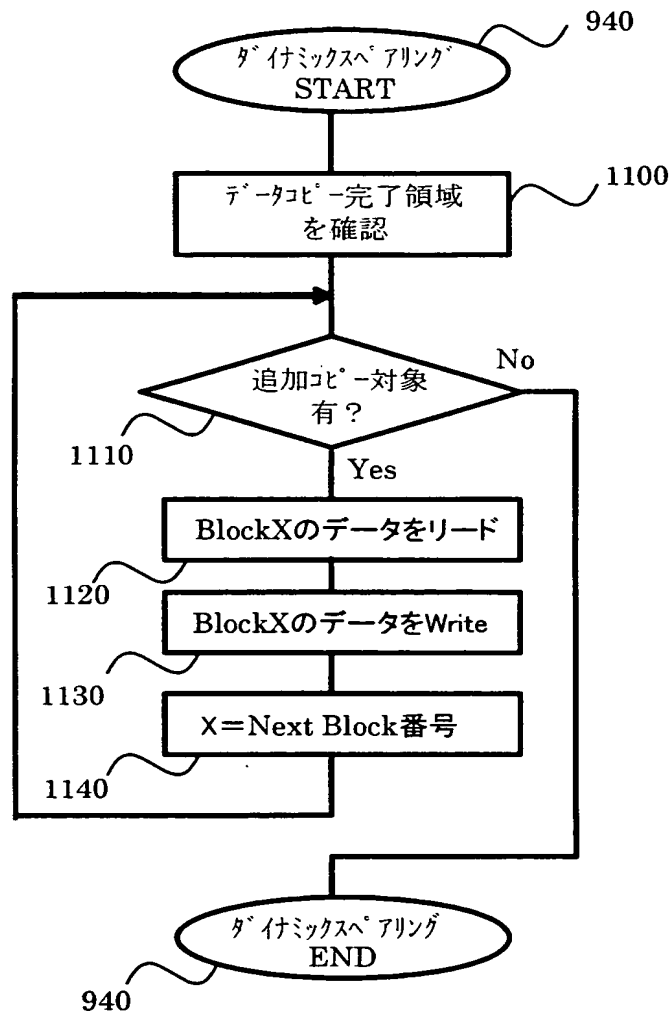
【図 5】



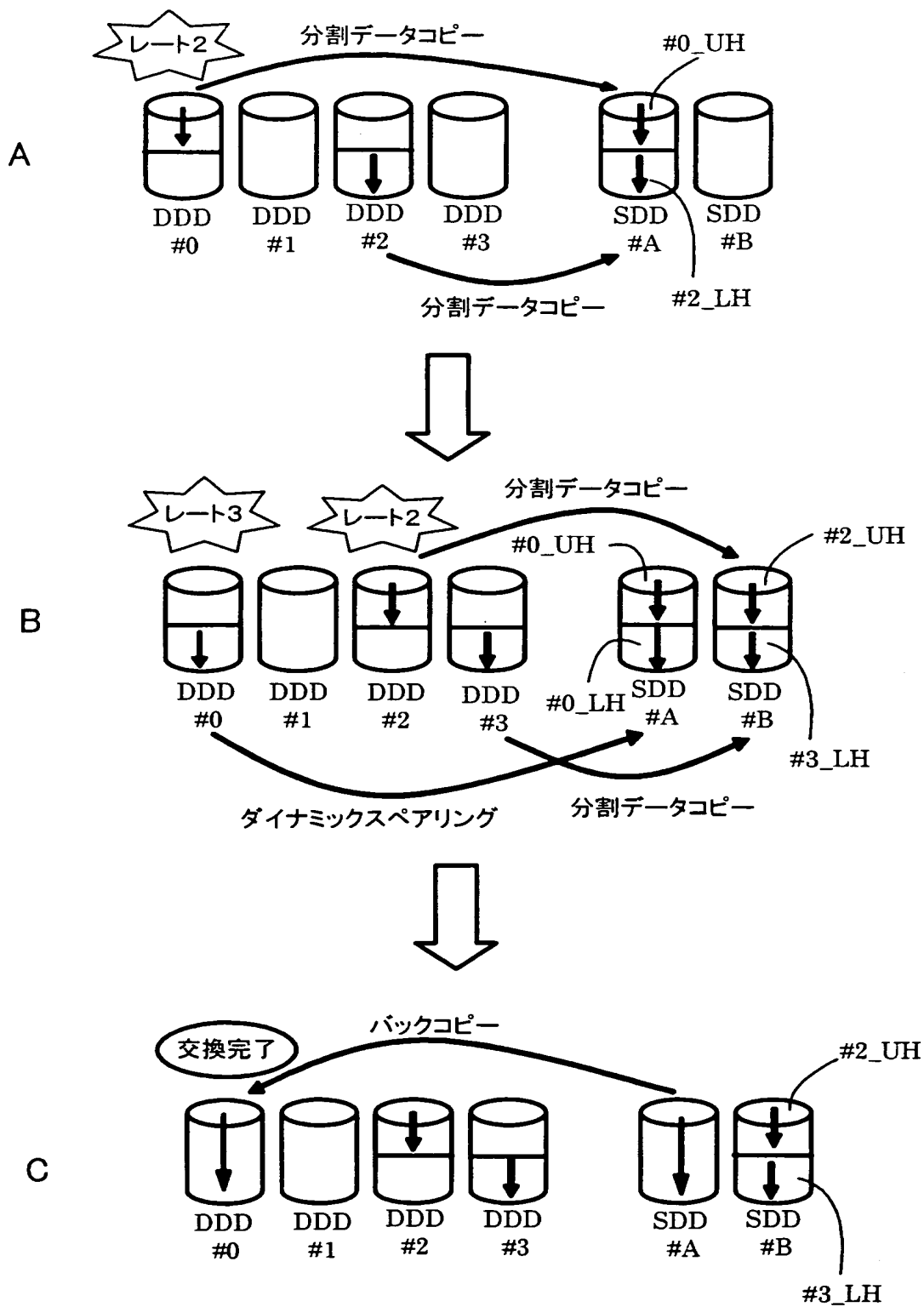
【図 6】



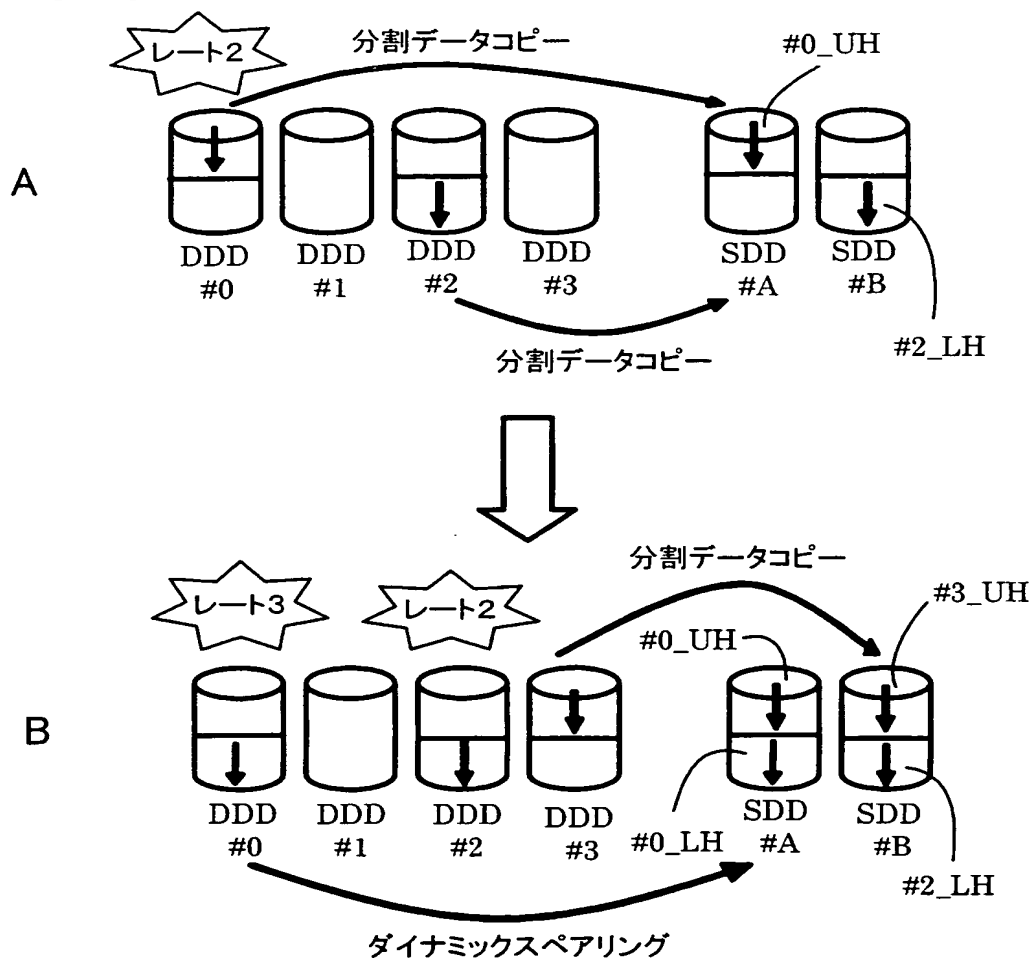
【図 7】



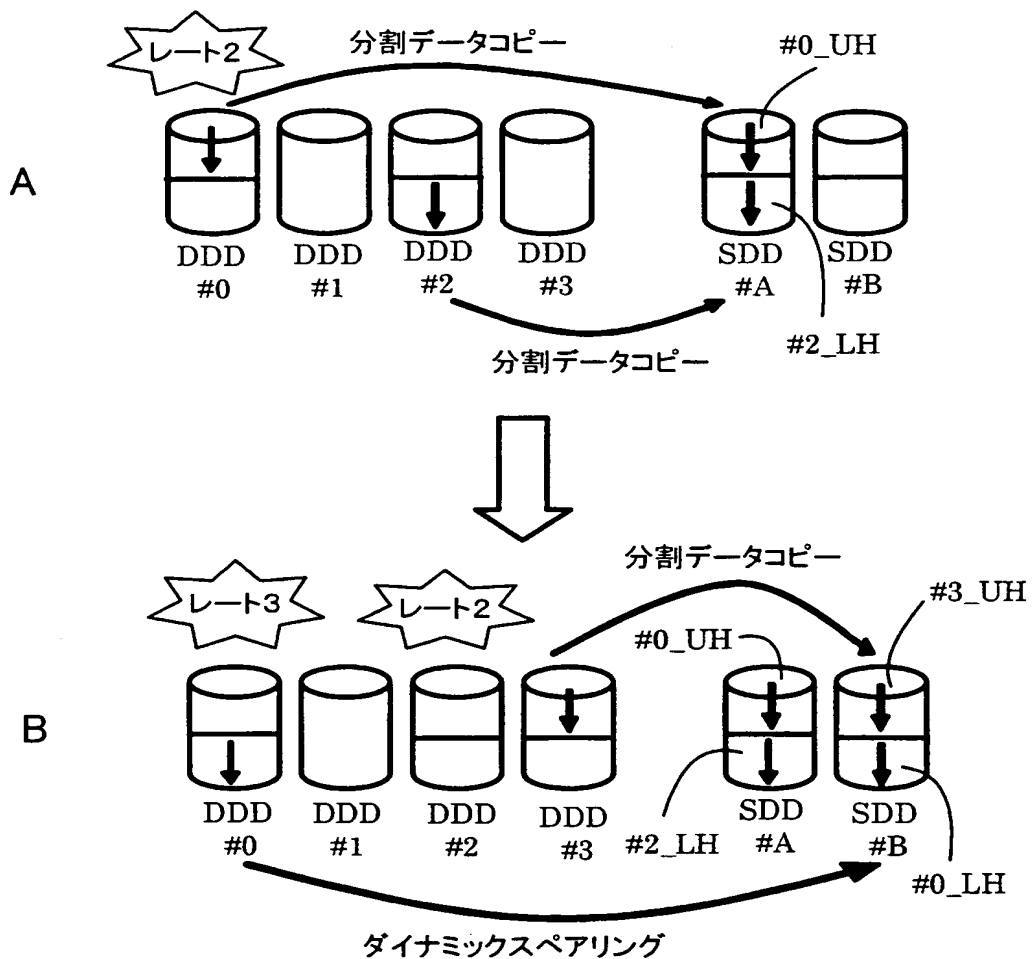
【図 8】



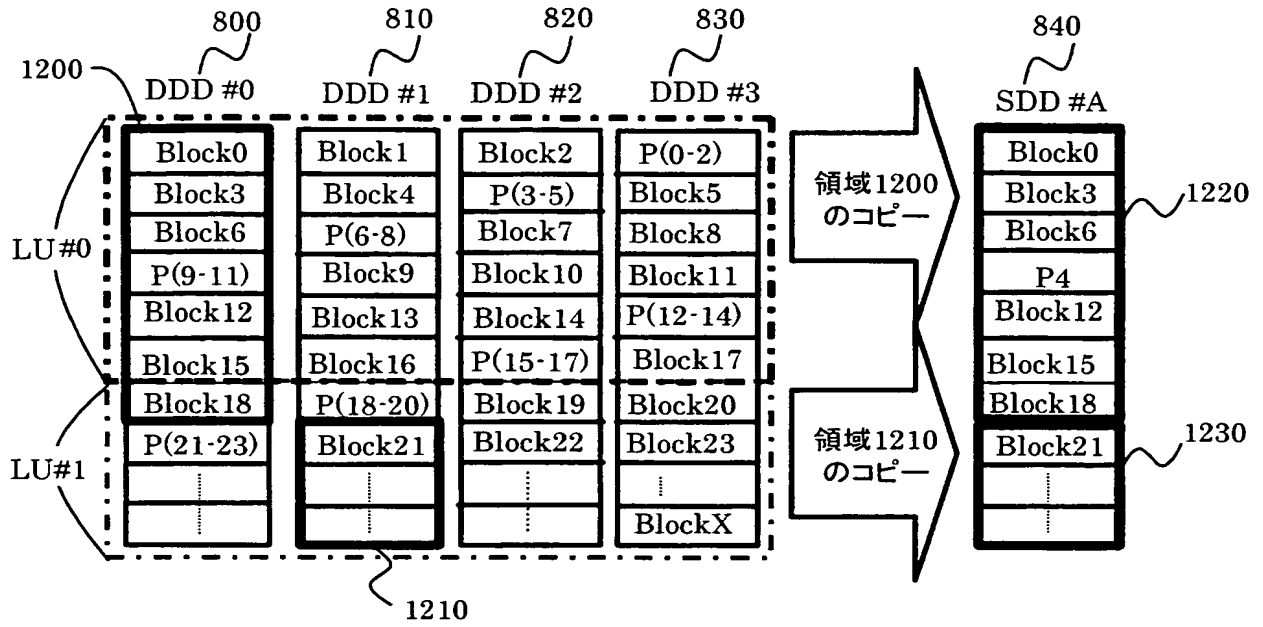
【図 9】



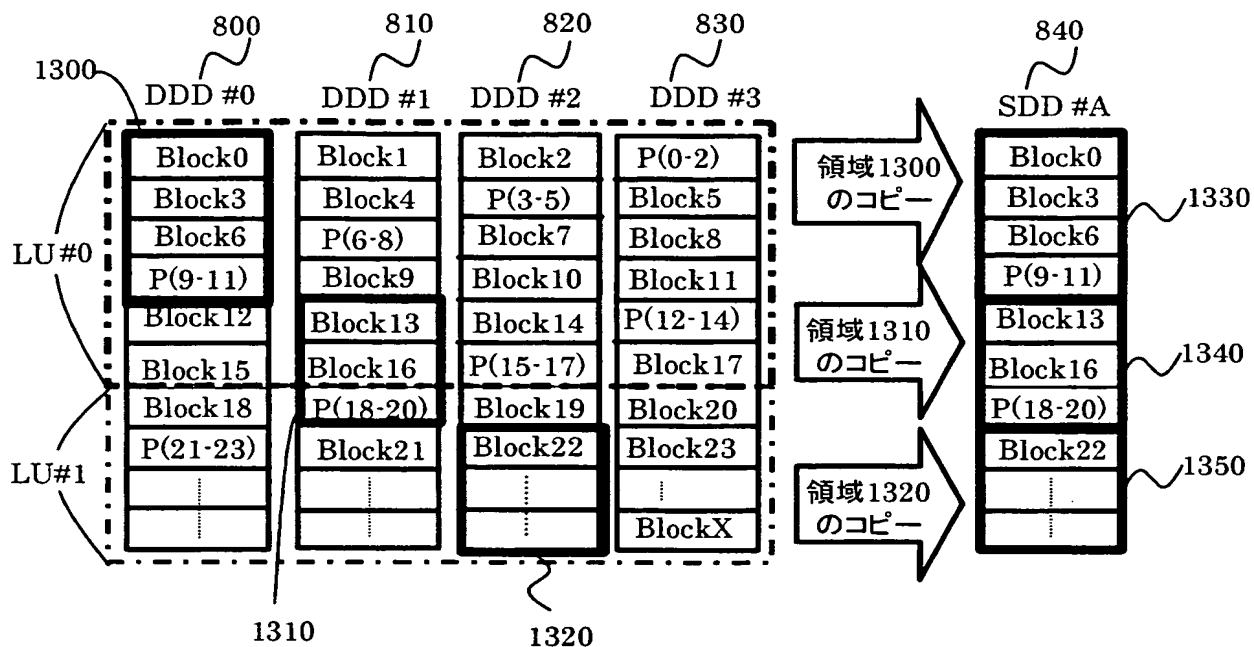
【図10】



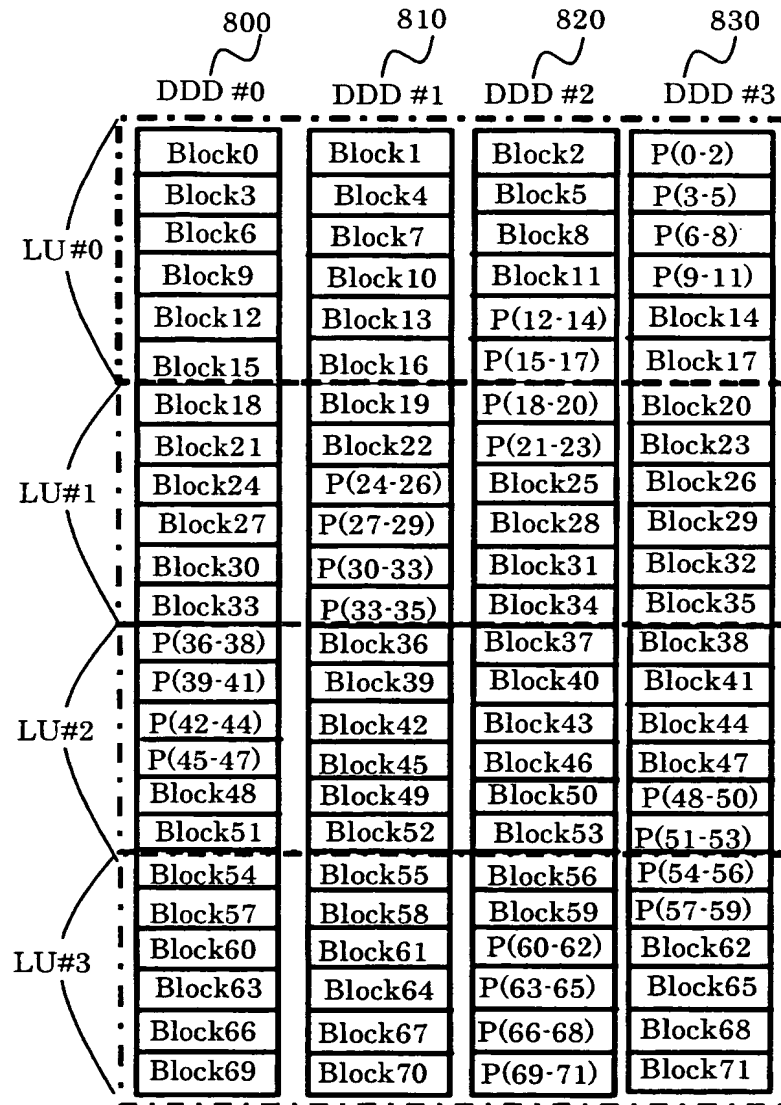
【図 11】



【図 1 2】



【図13】



【書類名】 要約書**【要約】**

【課題】 スペアディスクを持つRAID構成のディスクアレイにおいて多重ディスク故障によるデータロストを防止する。

【解決手段】 或るパリティグループ内の個々のディスクドライブ#0～#3毎にエラー発生回数から故障発生の可能性を推測し、故障発生の可能性がある程度に高い2つのディスクドライブ#0、#2から、データストライプにて互いに異なる2つの半分サイズの記憶領域#0_UHと#2_LHをそれぞれ選び、スペアディスク#Aへコピーする（分割データコピー）。いずれか一方のディスクドライブ#0の故障発生の可能性が更に高まると、そのディスクドライブ#0の残りの半分の記録領域#0_LHをスペアディスク#Aへコピーする（ダイナミックスペアリング）。分割データコピーにより、ダイナミックスペアリング時間が短縮され、かつ多重ディスク故障によるデータロストの可能性が低減する。

【選択図】 図8

認定・付加情報

特許出願の番号	特願 2003-354557
受付番号	50301709400
書類名	特許願
担当官	第七担当上席 0096
作成日	平成15年10月16日

<認定情報・付加情報>

【提出日】 平成15年10月15日

特願 2 0 0 3 - 3 5 4 5 5 7

出 願 人 履 歴 情 報

識別番号 [0 0 0 0 0 5 1 0 8]

1. 変更年月日	1 9 9 0 年 8 月 3 1 日
[変更理由]	新規登録
住 所	東京都千代田区神田駿河台 4 丁目 6 番地
氏 名	株式会社日立製作所